UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

SINGULAR COMPUTING LLC,

Plaintiff,

v.

GOOGLE LLC,

Defendant.

Civil Action No. 1:21-cv-12110

JURY TRIAL DEMANDED

COMPLAINT FOR PATENT INFRINGEMENT

Plaintiff, Singular Computing LLC ("Singular"), for its complaint against defendant, Google LLC ("Google"), alleges as follows:

THE PARTIES

1. Singular is a Delaware limited liability company having its principal places of business at 10 Regent Street, Newton, Massachusetts 02465 and The Cambridge Innovation Center, 1 Broadway, Cambridge, Massachusetts 02142.

2. Google is a Delaware limited liability company having regular and established places of business in this District, including a major office complex in Cambridge, Massachusetts with over 1,500 employees. Google may be served with process through its registered agent, Corporation Service Company, 84 State Street, Boston, Massachusetts 02109.

JURISDICTION

3. This is a civil action for patent infringement under the patent laws of the United States, 35 U.S.C. §§ 271, *et seq*. This Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 and 1338(a).

4. This Court has general personal jurisdiction over Google because Google is engaged in substantial and continuous activity, which is not isolated, at its regular and

established places of business within this judicial district. This Court has specific personal jurisdiction over Google because Google has also committed acts of infringement within this judicial district giving rise to this action and has established more than minimum contacts within this judicial district such that the exercise of jurisdiction over Google by this Court would not offend traditional notions of fair play and substantial justice.

5. Venue is proper in this judicial district pursuant to 28 U.S.C. §§ 1391(b)-(c) and 1400(b) because Google maintains regular and established places of business and has committed acts of patent infringement within this judicial district.

FACTUAL BACKGROUND

6. Singular was founded by Dr. Joseph Bates to, *inter alia*, design, develop, and produce computers having new architectures, including the patented computer architectures at issue in this case. Dr. Bates is the President and Chief Executive Officer of Singular. Since 2009, Singular has continuously operated out of the Boston area.

7. Dr. Bates's interest in computer science dates back to at least 1969, when, at the age of thirteen, he was admitted to Johns Hopkins University as an undergraduate. His success in college sparked a pilot program for exceptionally gifted youths, which led to the widely-recognized Johns Hopkins Center for Talented Youth (also known as "CTY"; *see* https://cty.jhu.edu) that has contributed to the intellectual development of over 165,000 academically advanced pre-college students, including Google founder Sergey Brin. By the age of 17, Dr. Bates had earned bachelor's and master's degrees from Johns Hopkins, both in the field of Computer Science. He continued his studies at Cornell University, where he earned his doctorate in Computer Science when he was 23 years old. Dr. Bates's research and teaching interests have centered around several cutting-edge computer science topics, including formal

logic, the design and implementation of computer programming languages, and artificial intelligence ("AI").

8. During his career working at the vanguard of computer science, Dr. Bates realized that, although the theoretical computing power inside computers (as represented by the number of transistors inside a computer) was growing exponentially under a phenomenon known as Moore's Law, the vast majority of that increase in computing power was not being made available to users. With then-existing computer architectures, even computers containing over a billion transistors were designed to typically perform only a handful of operations per unit of time ("period") when using CPUs. Such conventional computers of the time typically performed only a few hundred operations per period when using GPUs.

9. In the course of his work, Dr. Bates realized that existing computing architectures prevented computers from achieving their full potential. Computers perform computations using *transistors*, semiconductor devices that control the flow of electric current. For the last 50 years, due to advances in semiconductor technology, the number of transistors inside computer chips has grown at an exponential rate, doubling roughly every two years. Computer chips in the early 1970s contained just a few thousand transistors, while many similar chips used today have over 10 *billion* transistors. Dr. Bates recognized, however, that computing power (as measured by, *e.g.*, the number of computations a computer performs each second) had not increased at the same rate. Dr. Bates further recognized that computing power was lagging behind transistor count because a computer built using conventional architectures—even though it included more transistors—did not use its transistors efficiently.

10. Dr. Bates devised improved computer architectures that allow a computer to make more efficient use of its physical resources (*e.g.*, its transistors). The novel architectures

invented by Dr. Bates involve computer processors that contain processing elements purposely designed to perform low precision operations at high dynamic range. These architectures allow computers to use transistors more efficiently and have broad applicability to a wide variety of computing applications. In particular, Dr. Bates's inventions have revolutionized the field of AI and his patented architectures have vastly increased the speed and performance of computer processors when executing AI applications.

11. A key difference between conventional computer architectures and Dr. Bates's invention relates to a computer's performance of arithmetic operations such as multiplication. Using a conventional computer architecture, a typical multiplier circuit contains on the order of a hundred thousand transistors or more. A computer built using Dr. Bates's patented architecture, on the other hand, includes multiplier circuits that require a far smaller number of transistors, making it possible to include a very large number of them on a single chip, thereby increasing the number of multiplications per second the computer is able to perform. Indeed, a computer that uses Dr. Bates's invention can potentially perform hundreds of times more multiplications per second than a conventional computer with the same number of transistors.

12. In some embodiments of Singular's patented novel computer architectures, processing elements that operate at low precision can be deployed within a computer in parallel configurations to further amplify their relatively higher efficiency. In still other exemplary configurations, large numbers of these processing elements can be deployed in conjunction with far smaller numbers of higher-precision processing elements found within conventional computer architectures.

13. Singular's revolutionary approach to computer architecture is described in a provisional patent application entitled "Massively Parallel Processing with Compact Arithmetic Element" that was filed in June of 2009 and made public in June of 2010.

14. After Dr. Bates filed this provisional patent application, he built a prototype computer using its novel architecture. The Singular prototype was able to execute a software program that, for example, was able to perform neural network image classification thirty (30) times faster than a conventional computer having comparable physical characteristics in terms of its number of transistors, its semiconductor fabrication process and its power draw.

15. As Singular was building prototypes of its new computer, Google belatedly recognized the limitations of its conventional computer architectures in providing users with computer-based services such as Translate, Photos, Search, Assistant, and Gmail. According to Google, these limitations caused a "scary and daunting" situation for Google. The situation arose as Google was starting to deliver these computer-based services by running AI software programs on its conventional computers. The situation was "scary and daunting" because the new AI software programs required far more computer operations per period than the software programs Google was previously executing to deliver such services. For example, by Google's own estimation, applying its new AI software programs to speech recognition services alone (e.g., Translate and Assistant) would increase the required number of operations per period so drastically that Google would have to at least double its total computing footprint.

16. Google realized it needed to use Dr. Bates's computer architectures to increase the number of computer operations per period executed by its computers. To this end, it has been using the accused TPUv2, v3 and v4 devices (together "accused TPUs") to deliver services such as Translate, Photos, Search, Assistant, Cloud and Gmail to the public. Google drives the

public's use of these services to enhance its Ads platform which, in turn, generates at least tens of billions of dollars per year in profit for Google. *See* https://www.statista.com/statistics/266206/googles-annual-global-revenue/.

17. As of 2017, Google housed its TPU computers in the United States in at least eight data centers. As of 2017, the approximate cost to build each data center was at least $1.5 billion. As Google recognized, unless it incorporated Dr. Bates's patented technology, it would have had to at least double its number of data centers in the U.S. to sixteen. Assuming a cost of $1.5 billion per new data center, this would have cost Google a total of at least $12 billion.

18. With the steep growth of its business since 2017, Google now maintains at least fourteen data centers in the United States for its TPU computers. *See* www.google.com/about/datacenters/locations/. The accused TPUs are installed and operated by Google in one or more of Google's data centers located at: Berkeley County, South Carolina; Council Bluffs, Iowa; The Dalles, Oregon; Douglas County, Georgia; Henderson, Nevada; Jackson County, Alabama; Lenior, North Carolina; Loudoun County, Virginia; Mayes County, Oklahoma; Midlothian, Texas; Montgomery County, Tennessee; New Albany, Ohio; Papillon, Nebraska, and Storey County, Nevada.

19. In the Securities and Exchange Commission Form 10-K filed by Google's parent Alphabet, Inc. ("Alphabet") for the fiscal year ending December 31, 2020, Alphabet reported net income of approximately $40.2 billion for 2020 revenues in excess of $182 billion.

THE PATENTS-IN-SUIT

20. On August 25, 2020, the United States Patent and Trademark Office ("USPTO") issued United States Patent No. 10,754,616, titled PROCESSING WITH COMPACT

ARITHMETIC PROCESSING ELEMENT ("the '616 patent"). The '616 patent is valid and enforceable.

21. On November 9, 2021, the USPTO issued United States Patent No. 11,169,775, titled PROCESSING WITH COMPACT ARITHMETIC PROCESSING ELEMENT ("the '775 patent"). The '775 patent is valid and enforceable.

22. The application to which the '616 patent and the '775 patent claim priority (No. 61/218,691) was filed on June 19, 2009.

23. Singular is the owner and assignee of all rights, title and interest in and to the '616 patent and the '775 patent, and holds all substantial rights therein, including the rights to grant licenses, to exclude others, and to enforce and recover past damages for infringement.

24. The claims asserted in this action are eligible for patenting under 35 U.S.C. § 101.

25. Claim 10 of the '616 patent recites the following limitations, each of which is found in the accused TPUs as set forth below:[1]

10. A computing system, comprising:

a host computer;

a computing chip comprising:

> *a processing element array comprising a plurality of first processing elements, wherein the plurality of first processing elements is no less than 5000 in number, wherein each of a first subset of the plurality of first processing elements is positioned at a first edge of the processing element array, and wherein each of a second subset of the plurality of first processing elements is positioned in the interior of the processing element array;*
>
> *an input-output unit connected to each of the first subset of the plurality of first processing elements;*

[1] Claim 10 of the '616 patent is a dependent claim; it depends from claim 8, which in turn depends from independent claim 7. It has been written herein in independent form, to include the limitations of claims 7 and 8 from which it depends.

> > *a plurality of processing element connections, each processing element connection connecting one of the plurality of first processing elements with another of the plurality of first processing elements, wherein each of the plurality of first processing elements is connected to at least one other of the plurality of first processing elements by at least one of the plurality of processing element connections;*
> >
> > *a plurality of memory units, wherein each of the plurality of first processing elements is associated with a corresponding one of the plurality of memory units, and wherein each of the plurality of memory units is local to its associated one of the plurality of first processing elements; and,*
> >
> > *a plurality of arithmetic units, wherein each of the plurality of first processing elements has positioned therein at least one of the plurality of arithmetic units; and,*
>
> *a host connection at least partially connecting the input-output unit with the host computer;*
>
> *wherein the plurality of arithmetic units each comprises a first corresponding multiplier circuit adapted to receive as a first input to the first corresponding multiplier circuit a first floating point value having a first binary mantissa of width no more than 11 bits and a first binary exponent of width at least 6 bits, and to receive as a second input to the first corresponding multiplier circuit a second floating point value having a second binary mantissa of width no more than 11 bits and a second binary exponent of width at least 6 bits;*
>
> *wherein the computing chip further comprises a plurality of second processing elements, wherein the plurality of second processing elements each comprises a second corresponding multiplier circuit adapted to receive as inputs to the second corresponding multiplier circuit two floating point values each of width at least 32 bits;*
>
> *wherein, other than the plurality of second processing elements, the computing chip has no other processing element that comprises a multiplier circuit adapted to receive as inputs to the multiplier circuit two floating point values each of width at least 32 bits;*
>
> *wherein the plurality of first processing elements is greater in number, by at least 100, than the plurality of second processing elements; and*
>
> *wherein said host computer is programmed to provide instructions to said computing chip that, when executed, cause said processing element array to perform an operation whose output is used to identify at least one image, from a plurality of images to be searched, that is similar to at least one input image.*

26. Claim 1 of the ’775 patent recites the following limitations, each of which is likewise found in the accused TPUs as set forth below:

> *1. A computing system, comprising:*

a host computer;

a computing chip comprising:

> *a processing element array comprising a first edge processing element positioned at a first edge of the processing element array, a second edge processing element positioned at the first edge of the processing element array, a first interior processing element positioned at a first location in the interior of the processing element array, and a second interior processing element positioned at a second location in the interior of the processing element array;*
>
> *a first processing element connection connecting the first edge processing element with the first interior processing element;*
>
> *a second processing element connection connecting the second edge processing element with the second interior processing element;*
>
> *an input-output unit connected to the first edge processing element and the second edge processing element;*
>
> *a first memory local to the first edge processing element;*
>
> *a second memory local to the second edge processing element;*
>
> *a third memory local to the first interior processing element;*
>
> *a fourth memory local to the second interior processing element; and,*
>
> *a fifth arithmetic unit;*
>
> *wherein the first edge processing element comprises a first arithmetic unit;*
>
> *wherein the second edge processing element comprises a second arithmetic unit;*
>
> *wherein the first interior processing element comprises a third arithmetic unit; and*
>
> *wherein the second interior processing element comprises a fourth arithmetic unit; and,*

a host connection at least partially connecting the input-output unit with the host computer;

wherein the first, second, third and fourth arithmetic units each comprises a corresponding multiplier circuit adapted to receive as a first input to the corresponding multiplier circuit a first floating point value having a first binary mantissa of width no more than 11 bits and a first binary exponent of width at least 6 bits, and to receive as a second input to the corresponding multiplier circuit a second floating point value having a second binary mantissa of width no more than 11 bits and a second binary exponent of width at least 6 bits;

wherein the fifth arithmetic unit comprises a corresponding multiplier circuit adapted to receive as inputs to the corresponding multiplier circuit two floating point values each of width at least 32 bits;

> *wherein the multiplier circuit corresponding to the first arithmetic unit comprises a first plurality of transistors and has no other transistors, the multiplier circuit corresponding to the second arithmetic unit comprises a second plurality of transistors and has no other transistors, the multiplier circuit corresponding to the third arithmetic unit comprises a third plurality of transistors and has no other transistors, the multiplier circuit corresponding to the fourth arithmetic unit comprises a fourth plurality of transistors and has no other transistors, and the multiplier circuit corresponding to the fifth arithmetic unit comprises a fifth plurality of transistors; and,*
>
> *wherein the fifth plurality of transistors exceeds in number each of the first plurality of transistors, the second plurality of transistors, the third plurality of transistors, and the fourth plurality of transistors.*

27. The inventions recited in claim 10 of the ’616 patent and claim 1 of the ’775 patent (together “the Asserted Claims”) were not conventional. Reducing the claimed inventions to practice required the design and manufacture of a computer that was fundamentally different from prior art computers. Existing prior art computers did not practice the invention, nor could they be easily reconfigured or modified to do so, because hardware of the time was unsuitable for implementing Dr. Bates’s inventions.

28. Computers built using the novel architecture of the Asserted Claims have many advantages over computers built using conventional architectures. These advantages include, but are not limited to, the following features:

a) including many more multiplier circuits on a single computer chip having a given set of resources, such as transistors, than prior art computer chips having a similar set of resources, by utilizing relatively imprecise multiplication circuits that require far fewer transistors than conventional, full-precision multiplication circuits;

b) performing a far greater number of operations per second—potentially on the order of 100 times or more—than a conventional computer of the time having the same number of transistors, semiconductor fabrication process and power draw; and

c) supporting software programs that require operations to be performed on numbers having high dynamic range.

29. Computers built using the novel architecture of the Asserted Claims have a relatively large number of lower-precision processing elements and a relatively smaller number of full-precision processing elements. This enables the claimed devices to support operations having a broader range of precisions. For example, the claims recite processing elements that are adapted to receive floating point values having a binary mantissa no more than 11 bits wide and an exponent at least 6 bits wide. As indicated above, the claims also recite a far smaller number of processing elements that receive floating point values that are at least 32 bits wide.

30. Collectively, the inventions recited in the Asserted Claims provide many advantages over the prior art. For example, the claimed systems use transistors more efficiently than those of the prior art, allowing a computer to perform on the order of 100 times or more operations per unit of time than a comparable prior art computer having the same number of transistors.

31. The Asserted Claims also address, *inter alia*, the inefficient use of transistors in prior art computer architectures described above. For example, as stated above, Dr. Bates's patented processing elements each utilize a smaller number of transistors than a full-precision processing elements of prior art computer architectures. This difference in the required number of transistors per processing element creates the opportunity to include more of the claimed low-precision processing elements in a computer, which in turn allows the computer to perform many more operations per second than a conventional computer having comparable computing resources (*e.g.*, number of transistors, power draw, etc.).

32. Dr. Bates's inventions solve the aforementioned problem of inefficient transistor usage with an unconventional and novel approach to computer architecture that is fundamentally different from prior art computer architectures. Dr. Bates's inventions were not obvious to one of ordinary skill in the art at the time of their invention. Prior art computer architectures lacked a large number of processing elements that operate at low precision but with high dynamic range. Before Dr. Bates invented it, such a computer did not exist, nor was it described in any patent or printed publication.

33. Indeed, when the priority application was filed in 2009, the novel architecture invented by Dr. Bates went against a general consensus among those of skill in the art that a computer with a large number of low-precision processing elements was incapable of acceptable performance. It was not obvious, and was in fact counterintuitive, to those skilled in the art as of 2009 to make a computer from a very large number of processing elements that operate at low precision and with high dynamic range, knowing that such a computer was going to be used by software programs to execute numerous tasks that each required hundreds, thousands or even millions of sequential arithmetic operations, with each such operation potentially producing errors that could accumulate over time as tasks were executed. Dr. Bates nonetheless conceived of, made and patented a working computer utilizing such low-precision processing elements, and demonstrated that such a computer could perform better than prior art computers across a variety of applications.

34. The Asserted Claims recite concrete structure for achieving more efficient computer functionality and are not directed to every way of achieving those results. The architecture described by the Asserted Claims departs from earlier approaches to solving problems with computer operations. The Asserted Claims are directed to specific structural

features that cause improvements in the capabilities of computing devices (*e.g.* using the computer's transistors more efficiently, thereby allowing software programs to perform more computational operations per second).

35. In short, Dr. Bates's fundamentally new, unconventional and novel approach to computer architecture was not obvious, conventional or routine to one of ordinary skill in the art at the time of the invention. For example, conventional architectures did not include large numbers of low-precision processing elements that operate at high-dynamic range; as a matter of fact, prior to Dr. Bates's invention thereof, such a computer did not exist.

36. Computer architects as of 2009 taught away from Dr. Bates's inventions. Dr. Bates's claimed low-precision processing elements each frequently generate, in response to a request to perform arithmetic operations on high dynamic range numbers, results that materially differ from the exact, accurate results of those operations. In 2009, it was counterintuitive to those of skill in the art to design a computer having processing elements that produce such intentionally imprecise results in executing millions of operations per second, wherein each such operation potentially produces errors that collectively can accumulate over time. Nonetheless, Dr. Bates conceived of and built a working computer that embodies the claimed invention and included a large number of low-precision processing elements.

37. The inventions claimed in the asserted patents ushered in a revolutionary increase in computer efficiency through improved architecture. The Asserted Claims recite architectural elements of computer design such as a relatively larger number of lower precision processing elements and a smaller number of higher precision processing elements, that all support high dynamic range. This difference in the number of processing elements creates the opportunity to pack into a computer having a normal number of transistors (*e.g.*, several billion) a very large

number of processing elements that can collectively perform many times more operations per second than a computer built using conventional architectures.

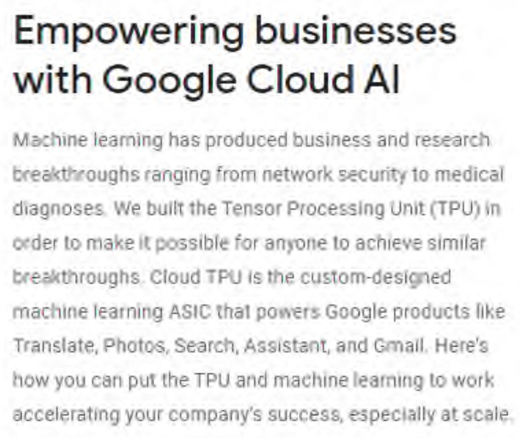
38. The inventions recited in the Asserted Claims were not conventional or routine as they provide more efficient use of a computer's transistors to perform an increased number of operations per period while supporting software programs that cause operations to be performed on numbers having high dynamic range and resulting in lower precision. Conventional computers, for example, even when designed for execution of AI software programs, did not include the concept of a computer based on such processing elements.

COUNT I
INFRINGEMENT OF THE '616 PATENT

39. Paragraphs 1-38 above are incorporated herein by reference.

40. As set forth below, Google has directly infringed, and continues to directly infringe, literally and/or by the doctrine of equivalents, at least claim and 10 of the '616 patent by making, testing, using, offering for sale, selling and/or importing into the United States the accused TPUs that are used with Google's existing data servers.

41. The accused TPUs power at least Google Translate, Photos, Search, Assistant and/or Gmail. For example, according to Google:



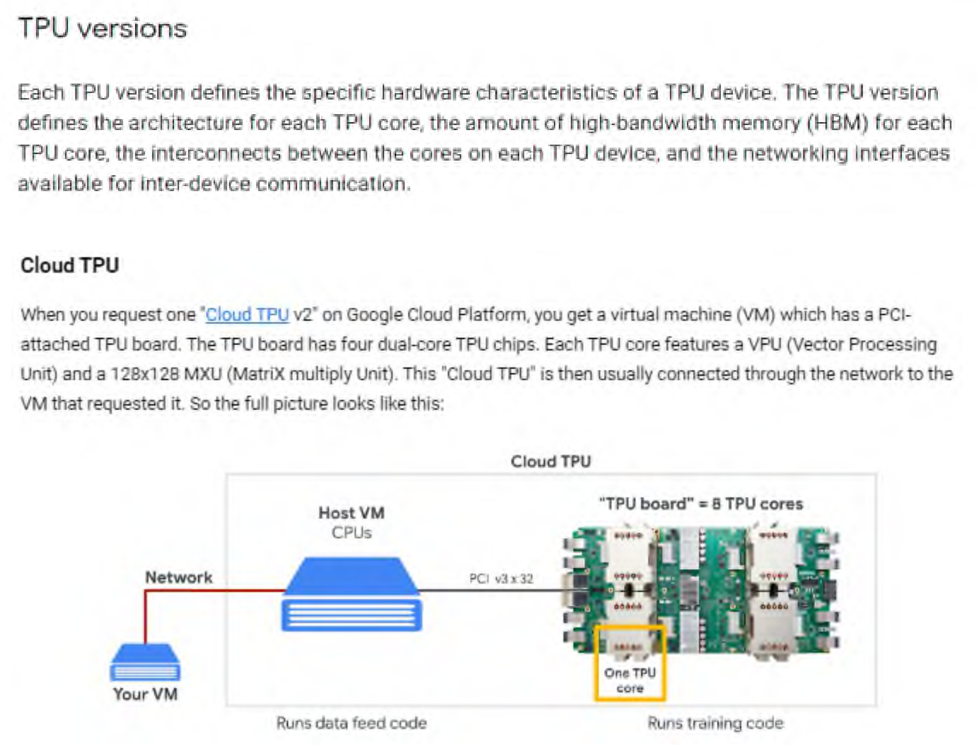
Empowering businesses with Google Cloud AI

Machine learning has produced business and research breakthroughs ranging from network security to medical diagnoses. We built the Tensor Processing Unit (TPU) in order to make it possible for anyone to achieve similar breakthroughs. Cloud TPU is the custom-designed machine learning ASIC that powers Google products like Translate, Photos, Search, Assistant, and Gmail. Here's how you can put the TPU and machine learning to work accelerating your company's success, especially at scale.

42. According to Google's Chief Executive Officer ("CEO"), Sundar Pichai, Google's TPU chips: have played a "big part" in Google's advances in AI services; are used "across all [Google's] products;" and are used "every time" a Google search is made. *See*, *e.g.*, blog.google/technology/developers/io21-helpful-google/.

43. According to cloud.google.com, "TPU v4 Pods are already widely deployed throughout Google data centers for [Google's] internal machine learning workloads and will be available via Google Cloud later this year."

44. Also, according to Google:



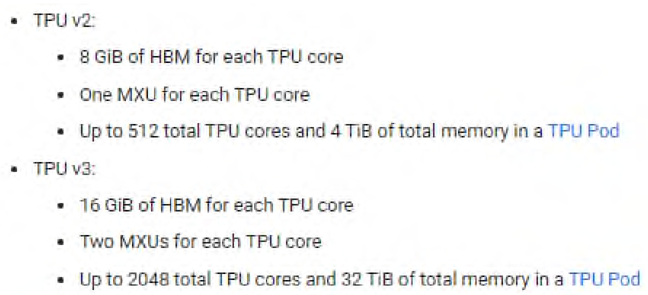
TPU versions

Each TPU version defines the specific hardware characteristics of a TPU device. The TPU version defines the architecture for each TPU core, the amount of high-bandwidth memory (HBM) for each TPU core, the interconnects between the cores on each TPU device, and the networking interfaces available for inter-device communication.

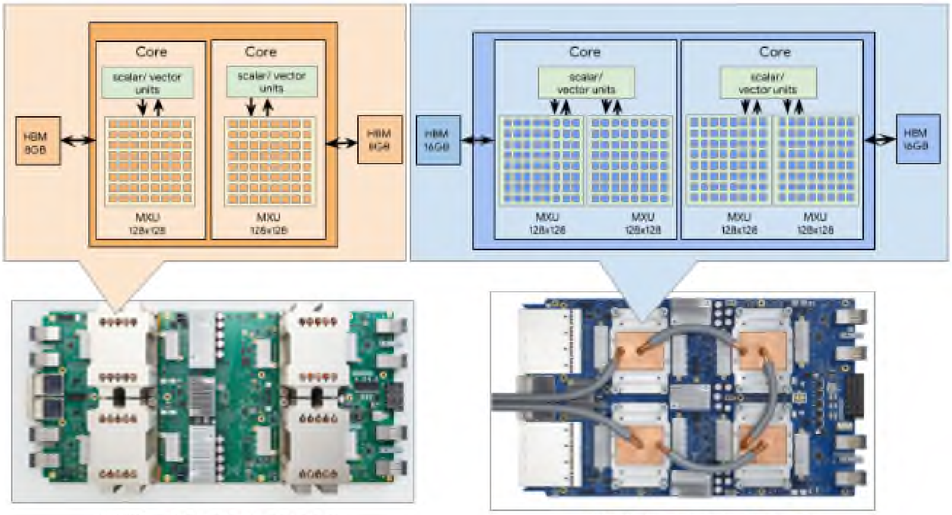
Cloud TPU

When you request one "Cloud TPU v2" on Google Cloud Platform, you get a virtual machine (VM) which has a PCI-attached TPU board. The TPU board has four dual-core TPU chips. Each TPU core features a VPU (Vector Processing Unit) and a 128x128 MXU (MatriX multiply Unit). This "Cloud TPU" is then usually connected through the network to the VM that requested it. So the full picture looks like this:

45. Google describes TPUs, *inter alia*, as follows:



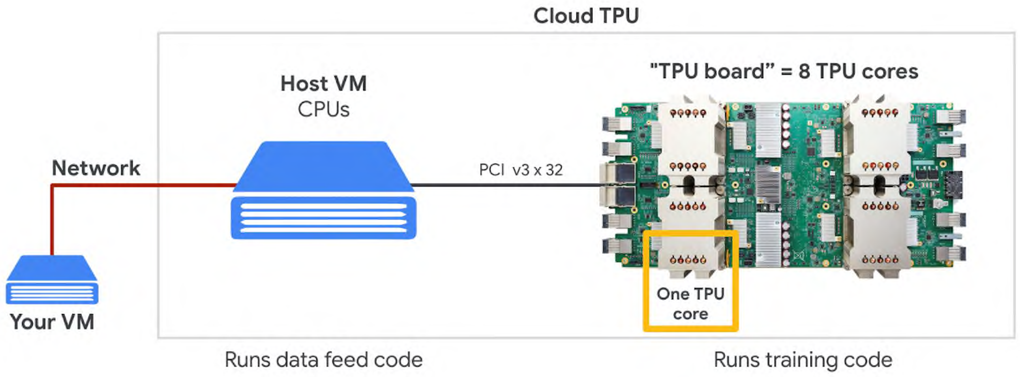
- TPU v2:
  - 8 GiB of HBM for each TPU core
  - One MXU for each TPU core
  - Up to 512 total TPU cores and 4 TiB of total memory in a TPU Pod
- TPU v3:
  - 16 GiB of HBM for each TPU core
  - Two MXUs for each TPU core
  - Up to 2048 total TPU cores and 32 TiB of total memory in a TPU Pod



TPU v2 - 4 chips, 2 cores per chip

TPU v3 - 4 chips, 2 cores per chip

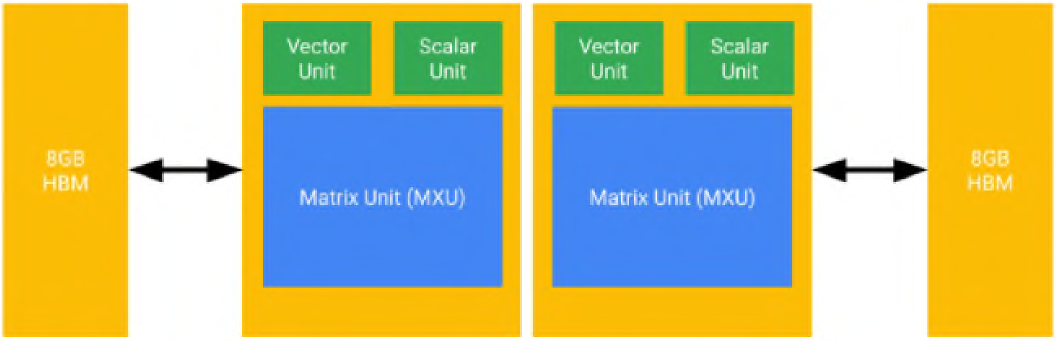
46. Google's Cloud Platform ("GCP") comprises a computer system having at least one input-output Host VM CPU connected to at least one TPU board having a plurality of TPU cores:



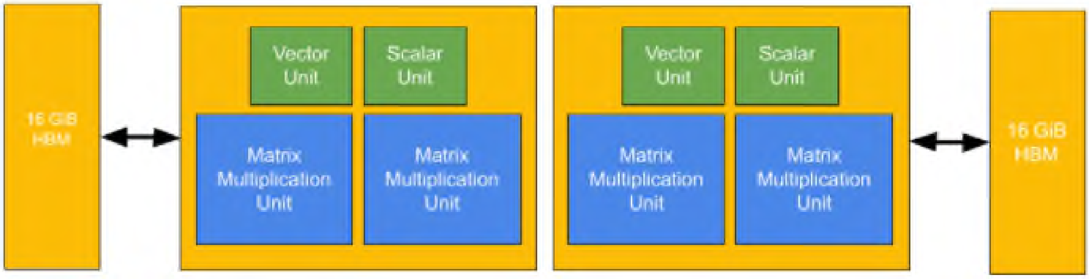
47. Each TPU board is connected to a Host VM CPU for loading and preprocessing data to feed to the TPUs. *See* https://cloud.google.com/tpu/docs/.

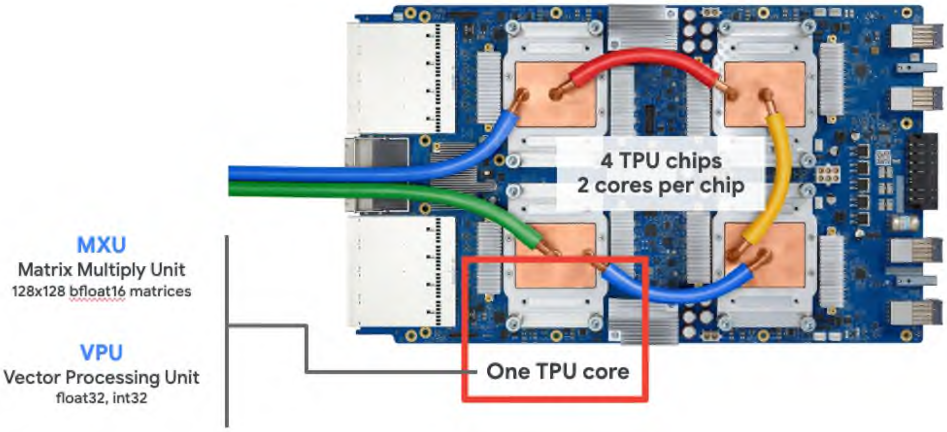
48. A TPU board includes a plurality of cores, each of which includes a MatriX Multiply Unit ("MXU") that runs matrix multiplications, a Vector Processing Unit ("VPU") and a Scalar Unit.

TPU v2:

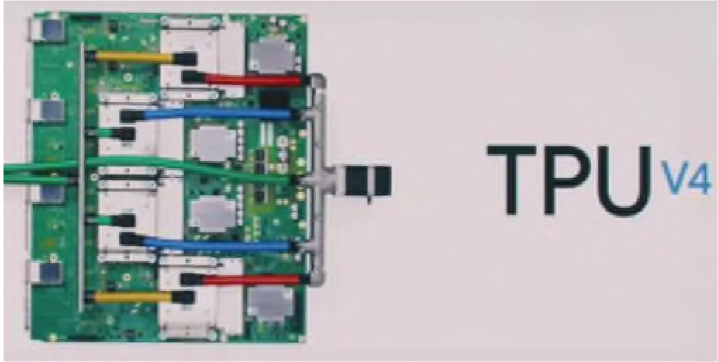


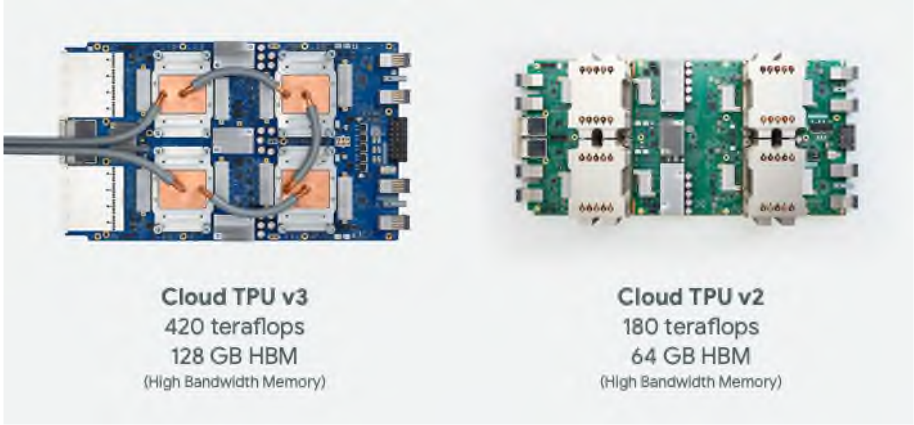
TPU v3:


16 GiB HBM
Vector Unit
Scalar Unit
Matrix Multiplication Unit
Matrix Multiplication Unit
Vector Unit
Scalar Unit
Matrix Multiplication Unit
Matrix Multiplication Unit
16 GiB HBM



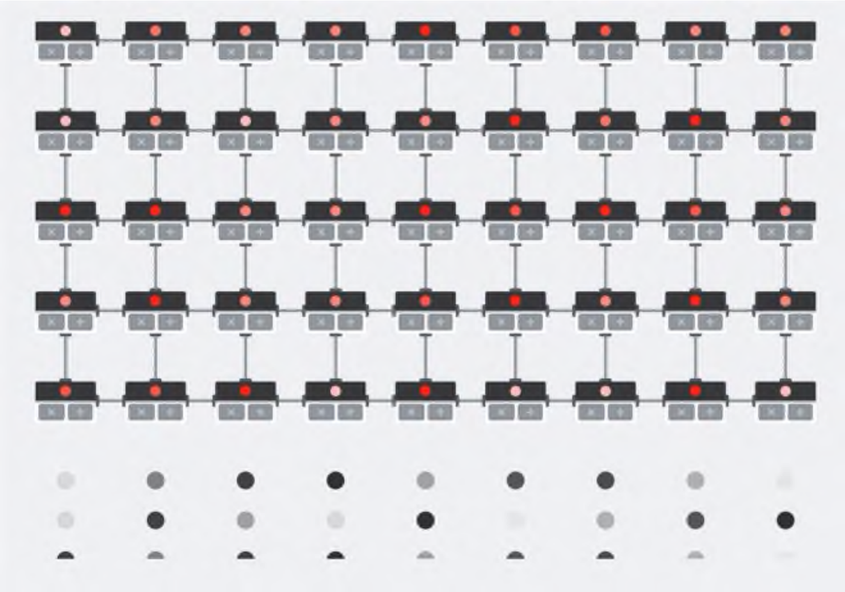
49. Each TPU chip on the board is external to the other three chips and includes at least one host interface to a host computer and High Bandwidth Memory ("HBM"):





50. TPU chips communicate with each other via high-bandwidth interconnects. *See* https://cloud.google.com/tpu/docs/.

51. Each TPU core includes at least one MXU that includes horizontally and vertically interconnected processing elements arranged in systolic arrays:

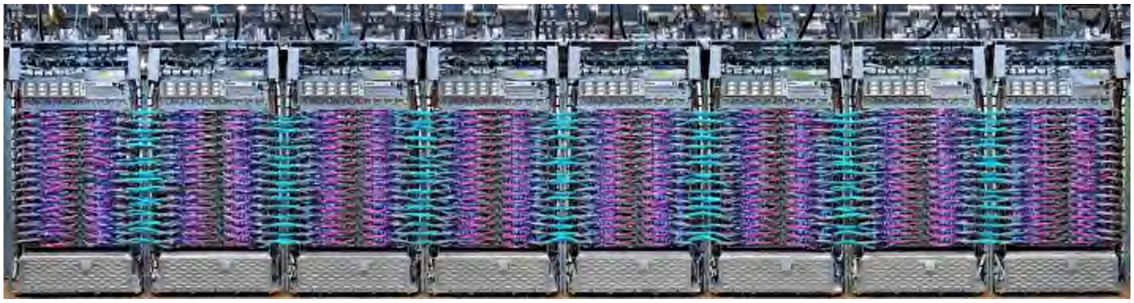


52. As shown above, a TPU systolic array includes at least five arithmetic processing elements, each comprising a plurality of transistors, in a column at the left edge of the array and a plurality of at least five connected adjacent processing elements interior and to the right of the at least five left edge processing elements.

53. As shown above, each processing element comprises an arithmetic unit that performs multiplication.

54. Each of the MXU processing elements has an associated memory. This memory is used, for example, to store "weights" or "parameters" as part of algorithms that relate to neural networks. *See*, *e.g.*, https://cloud.google.com/tpu/docs/beginners-guide ("the TPU loads the parameters from memory into the matrix of multipliers and adders").

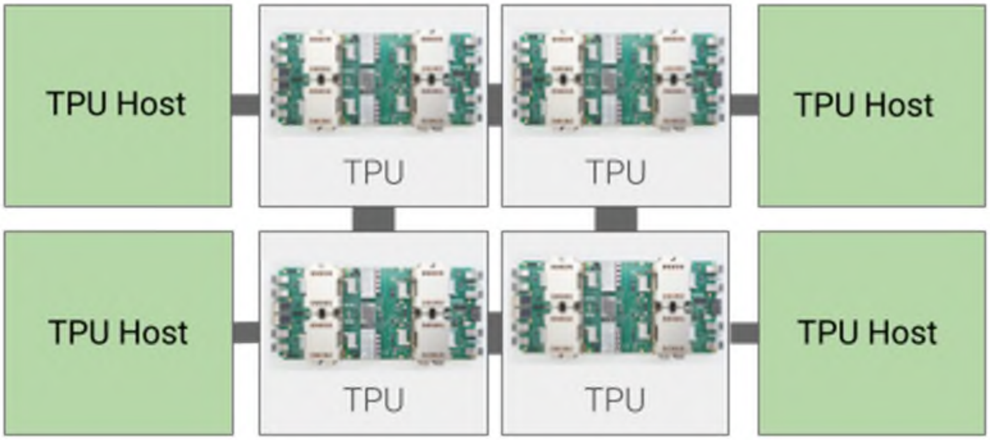
55. In Google's data centers, numerous TPU boards are connected together in slices and pods as shown, for example, below:



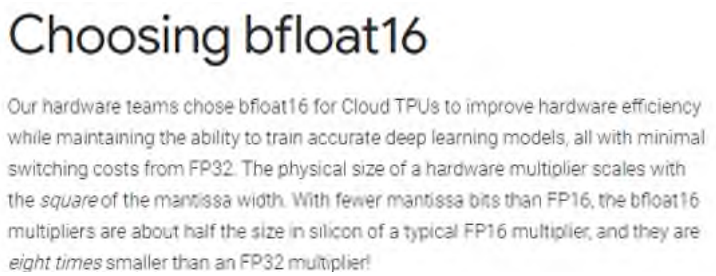
56. A TPU v3 pod may have up to 2,048 TPU cores and 32 TiB of memory. According to Google's CEO, Sundar Pichai, at Google I/O 2021, a TPU v4 Pod has 4,096 v4 chips and is capable of executing an exaflop of floating point operations per second.

57. A TPU pod includes at least five TPU boards.

58. In a TPU pod, a TPU host is connected to a TPU board:



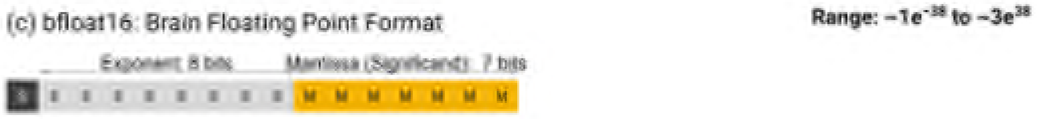
59. Each MXU processing element comprises a plurality of transistors that perform the operation of multiplication at bfloat16 precision.



Choosing bfloat16

Our hardware teams chose bfloat16 for Cloud TPUs to improve hardware efficiency while maintaining the ability to train accurate deep learning models, all with minimal switching costs from FP32. The physical size of a hardware multiplier scales with the *square* of the mantissa width. With fewer mantissa bits than FP16, the bfloat16 multipliers are about half the size in silicon of a typical FP16 multiplier, and they are *eight times* smaller than an FP32 multiplier!

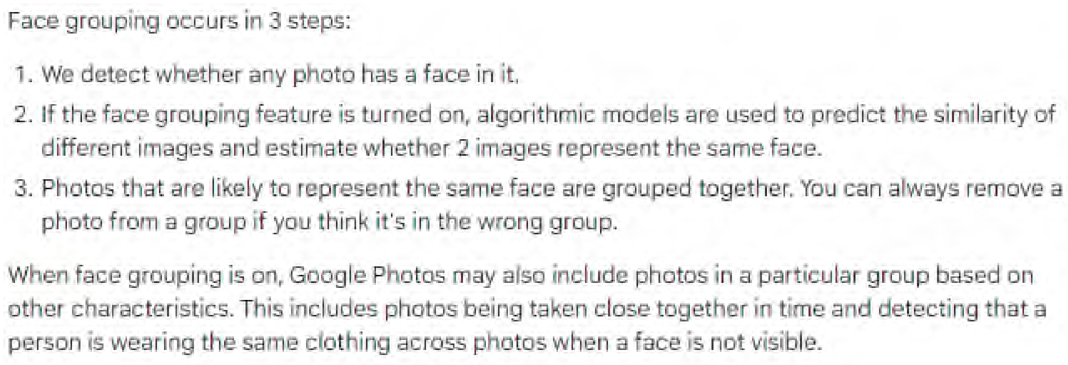
60. The bfloat16 used in the TPUs has a sign bit, 8 exponent bits and 7 mantissa bits:


(c) bfloat16: Brain Floating Point Format
Range: $-1e^{-38}$ to $-3e^{38}$
Exponent: 8 bits
Mantissa (Significand): 7 bits
S
E E E E E E E E
M M M M M M M

61. As described by Google above, bfloat16 utilizes a binary mantissa that is no more than 11 bits and a binary exponent of at least 6 bits. Google copied the idea to use an array of processing elements that perform floating-point arithmetic using a low-precision, high dynamic range number format from Dr. Bates.

62. The accused TPUs perform prediction (also called “inference”) and training. TPU v2 was the first version of Google’s TPU products that used floating point calculations to perform prediction and training. Google’s first TPU product, the TPU v1 (not accused in this complaint) was based on integer arithmetic and was unable to perform floating-point operations. The TPU v1 did not perform training and was used by Google for inference.

63. For example, the accused TPUs perform prediction for users of Google Photos to analyze the similarity of a user-input image to other images searched by Google to estimate whether two images are similar and may represent the same face:



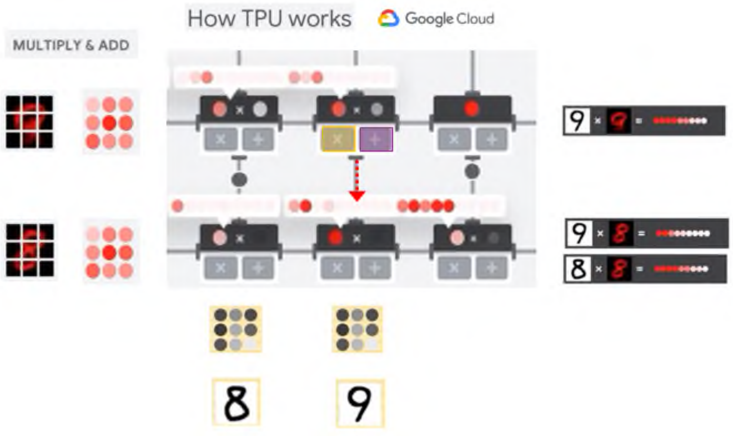
Face grouping occurs in 3 steps:

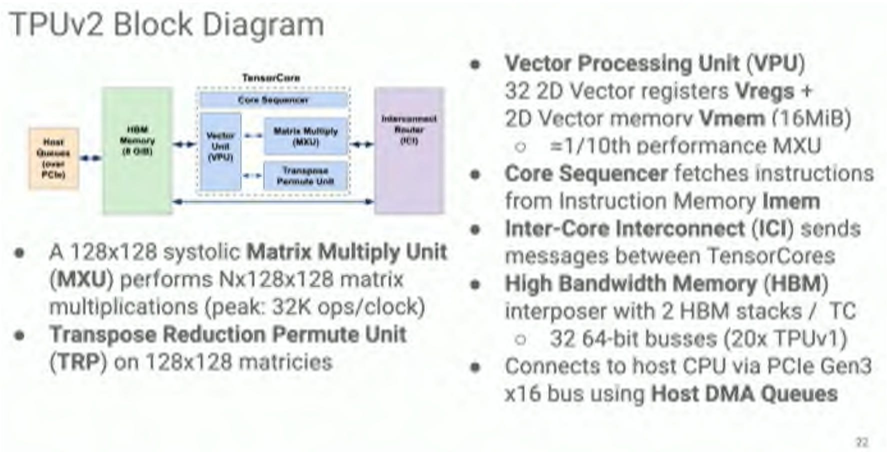
1. We detect whether any photo has a face in it.
2. If the face grouping feature is turned on, algorithmic models are used to predict the similarity of different images and estimate whether 2 images represent the same face.
3. Photos that are likely to represent the same face are grouped together. You can always remove a photo from a group if you think it's in the wrong group.

When face grouping is on, Google Photos may also include photos in a particular group based on other characteristics. This includes photos being taken close together in time and detecting that a person is wearing the same clothing across photos when a face is not visible.

https://support.google.com/photos/. The accused TPUs perform a similar predictive image comparison functionality for users of, *inter alia*, Google Lens (*see* https:/lens.google/ howlensworks/) and Google Images (*see* https://images.google.com). Google also offers its reverse image technology performed by the accused TPUs to third parties such as The New York Times and Box with instructions on how to use the technology with the intent that third parties use the technology in such an infringing manner. *See* https://cloud.google.com/vision.

64. According to Google, TPUs perform multiplication as follows:

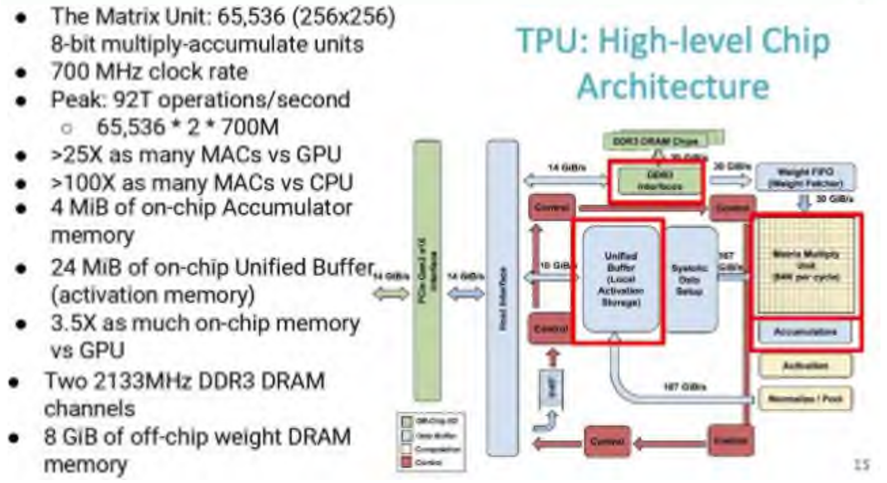




TPUv2 Block Diagram

- A 128x128 systolic **Matrix Multiply Unit** (**MXU**) performs Nx128x128 matrix multiplications (peak: 32K ops/clock)
- **Transpose Reduction Permute Unit** (**TRP**) on 128x128 matricies

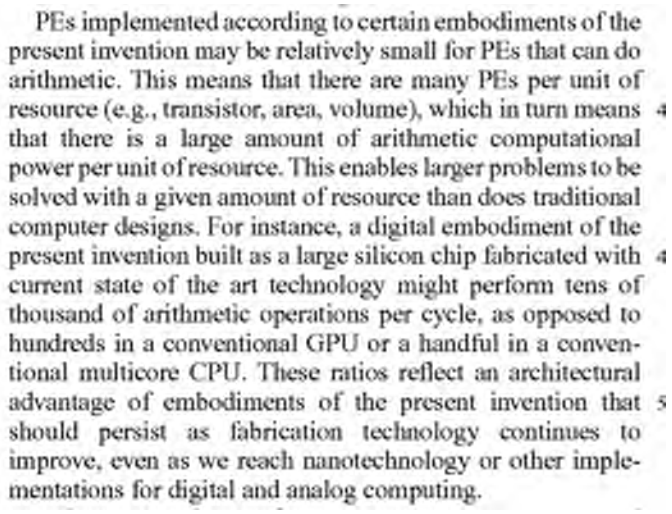
- **Vector Processing Unit** (**VPU**) 32 2D Vector registers **Vregs** + 2D Vector memory **Vmem** (16MiB)
  - ○ =1/10th performance MXU
- **Core Sequencer** fetches instructions from Instruction Memory **Imem**
- **Inter-Core Interconnect** (**ICI**) sends messages between TensorCores
- **High Bandwidth Memory** (**HBM**) interposer with 2 HBM stacks / TC
  - ○ 32 64-bit busses (20x TPUv1)
- Connects to host CPU via PCIe Gen3 x16 bus using **Host DMA Queues**



- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
- 700 MHz clock rate
- Peak: 92T operations/second
  - ○ 65,536 * 2 * 700M
- >25X as many MACs vs GPU
- >100X as many MACs vs CPU
- 4 MiB of on-chip Accumulator memory
- 24 MiB of on-chip Unified Buffer (activation memory)
- 3.5X as much on-chip memory vs GPU
- Two 2133MHz DDR3 DRAM channels
- 8 GiB of off-chip weight DRAM memory

TPU: High-level Chip Architecture

65. In knowingly incorporating Dr. Bates's patented computer architectures into the accused TPUs, Google reaps the very same benefits that were predicted by Dr. Bates in his patent application more than 10 years ago. As predicted by Dr. Bates:



> PEs implemented according to certain embodiments of the present invention may be relatively small for PEs that can do arithmetic. This means that there are many PEs per unit of resource (e.g., transistor, area, volume), which in turn means that there is a large amount of arithmetic computational power per unit of resource. This enables larger problems to be solved with a given amount of resource than does traditional computer designs. For instance, a digital embodiment of the present invention built as a large silicon chip fabricated with current state of the art technology might perform tens of thousand of arithmetic operations per cycle, as opposed to hundreds in a conventional GPU or a handful in a conventional multicore CPU. These ratios reflect an architectural advantage of embodiments of the present invention that should persist as fabrication technology continues to improve, even as we reach nanotechnology or other implementations for digital and analog computing.

66. Google's infringement of the '616 patent is and has been willful.

67. Less than two years after the filing of his provisional application in June 2009, Dr. Bates and Google executed a Non-Disclosure Agreement ("NDA") prepared by Google. *See* Amended Answer (Dkt. No. 57) in case No. 1:19-cv-12551 (D. Mass.) ("Am. Ans."), ¶ 17.

68. On November 3, 2010, Joseph Bates forwarded a document titled "COMPUTING 10,000X MORE EFFICIENTLY" by email to Astro Teller at the email address astroteller@google.com.

69. On November 3, 2010, Joseph Bates discussed Singular's technology by telephone with Astro Teller while Astro Teller was in Massachusetts for, *inter alia*, a meeting with Joseph Bates at the Massachusetts Institute of Technology Media Lab.

70. On December 9, 2010, Joseph Bates forwarded a document titled "COMPUTING 10,000X MORE EFFICIENTLY" by email to Sebastian Thrun at the email address thrun@google.com with copies to Astro Teller and Sergio Gandara.

71. On December 9, 10 and 21, 2010, Joseph Bates discussed Singular's technology by telephone with Astro Teller.

72. After receiving the document titled "COMPUTING 10,000X MORE EFFICIENTLY," Astro Teller discussed Singular's technology with Larry Page and/or Sergey Brin.

73. On January 28, 2011, Joseph Bates discussed Singular's technology by telephone with Astro Teller.

74. On June 9, 2011, Joseph Bates discussed Singular's technology with Astro Teller, Sebastian Thrun and Megan Smith at a meeting at the Massachusetts Institute of Technology Media Lab.

75. On June 21, 2011, Joseph Bates forwarded a document titled "APPLICATIONS / MARKETS / AND DEALS" by email to Astro Teller at the email address astroteller@google.com.

76. On June 22, 2011, Joseph Bates forwarded a document titled "APPLICATIONS / MARKETS" by email to Astro Teller at the email address astroteller@google.com.

77. On June 24, 2011, Joseph Bates met with Astro Teller, Johnny Chen, and others from Google to discuss Singular's technology.

78. On June 22, 2011, Joseph Bates forwarded a document titled "SINGULAR COMPUTING" by email to Astro Teller at the email address astroteller@google.com.

79. On June 22, 2011, Joseph Bates forwarded a document titled "APPLICATIONS / MARKETS" by email to Astro Teller at the email address astroteller@google.com.

80. On September 17, 2013, Joseph Bates met with Google's Jeffrey Dean, Quoc Le and others at Google. Pursuant to the NDA between Google and Singular, a slide presentation titled "MULTI-MILLION CORE PROCESSORS AND THEIR APPLICATIONS" was loaded onto a Google laptop from which Dr. Bates displayed the slides to Dean and Le. Thereafter, on September 17, 2013, Dean emailed Dr. Bates stating: "A few folks here are interested in seeing if we can train neural nets with various kinds of computational inaccuracies."

81. On January 22, 2014, Dr. Bates emailed Jeffrey Dean referencing "Singular's hardware, software, patents, experience, etc." In an email response dated January 23, 2014, Jeffrey Dean stated, *inter alia*, that he had "passed this info along to two people I think are most relevant within Google".

82. On or around January 24, 2014, Dr. Bates forwarded a presentation titled "MANY-MILLION CORE PROCESSORS AND THEIR APPLICATIONS" to Nanette Boden

at the email address nanboden@google.com with copies to Jeff Dean and Norm Jouppi. The presentation stated that it was "Confidential, per Google/Singular MNDA, March 2011." In the presentation, Dr. Bates warned Google that Singular had patent protection relating to the disclosed Singular technology.

83. On February 2, 2017, Dr. Bates met with Astro Teller and Tammo Spalink at Google in Mountain View, California to make a presentation and demonstration of Singular's patented technology. On February 27, 2017, James Laudon asked Dr. Bates for a set of the presentation slides. On March 1, 2017, Dr. Bates sent a copy of the slides, titled "APPROXIMATE COMPUTING, EMBEDDED AI, BILLION CORE SYSTEMS," to James Laudon.

84. On February 20, 2017, Obi Felten of Google's X team informed Dr. Bates by email that "Catherine Tornabene from the X IP legal team . . . will review your patent family."

85. On March 1, 2017, Jenn Wall, then a commercial lawyer in Google's X team, forwarded to Dr. Bates by email a draft Mutual Confidentiality and Non-Disclosure Agreement ("draft MNDA"). Paragraph 8 of the draft MNDA contained the following language:

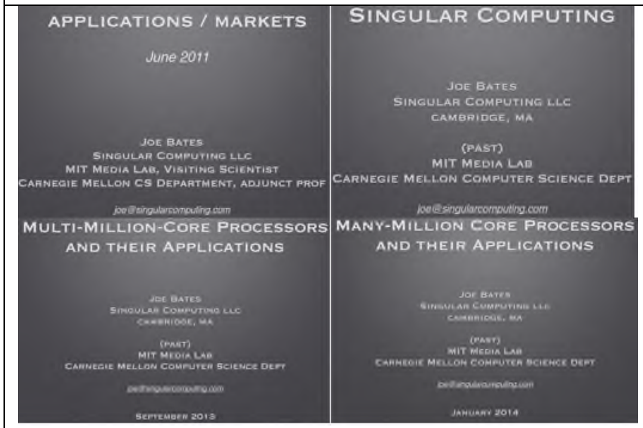
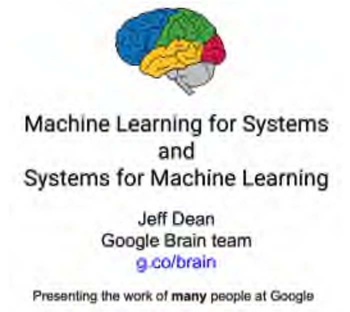
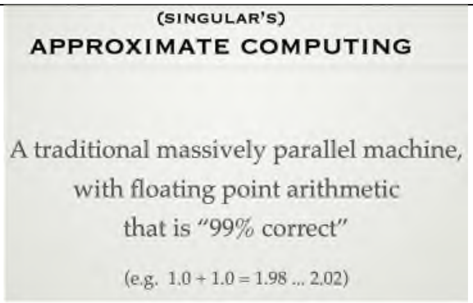
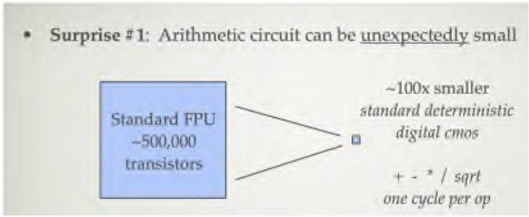
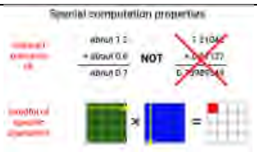
> The Company [Singular] waives any right to allege willful infringement based on notice to or knowledge by Google of any patent identified by the Company to Google (a) under this Agreement or (b) in any communication related to the Transaction prior to the effective date of this Agreement.

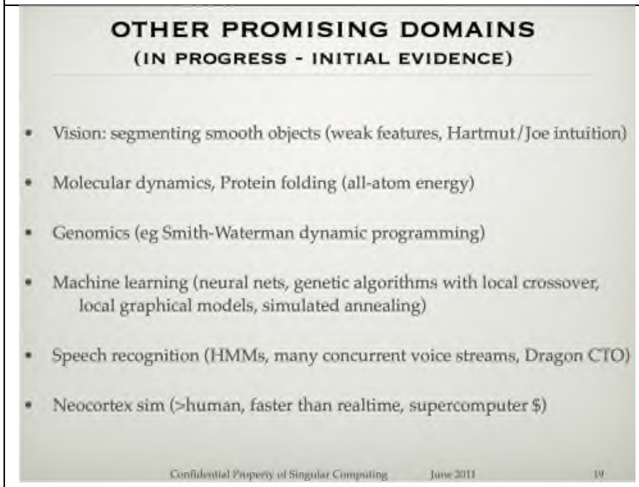
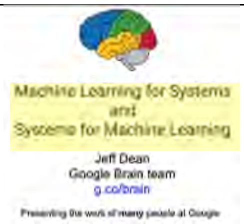
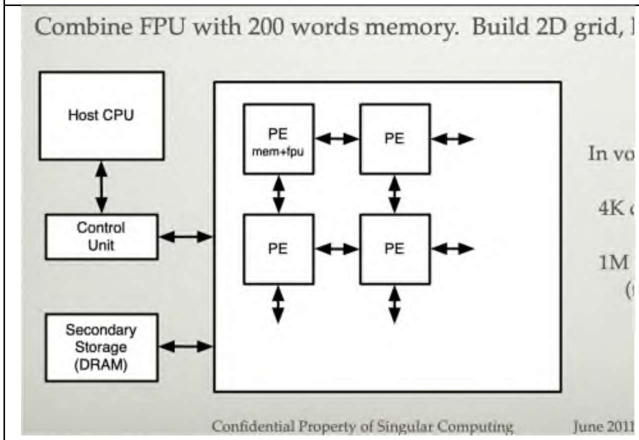
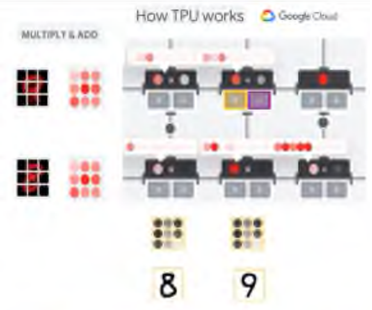
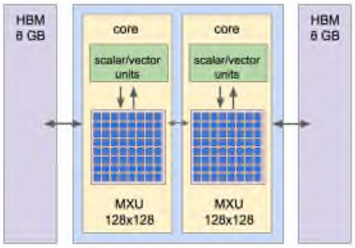
86. Dr. Bates did not sign Google's draft MNDA that was forwarded on March 1, 2017.

87. During the course of these meetings, Dr. Bates discussed his computer theories with certain Google employees. *See* Am. Ans., ¶ 19. Dr. Bates also advised Google that his disclosed computer architecture and S1 prototype were patent-protected. For example, in a presentation titled "Approximate Computing, Embedded AI, Billion Core Systems," Dr. Bates

informed representatives of Google in February 2017 that Singular had "broad patents granted in U.S. and internationally."

88. Following Dr. Bates's disclosure of his inventions, Google copied and adopted Dr. Bates's patented inventions, incorporating the disclosed technology into the accused TPUs installed in Google's data centers. This is apparent from a comparison of Dr. Bates's patented architecture and that of the accused TPUs. It is also apparent from an exemplary comparison of the disclosures made in writing by Dr. Bates to Google with the properties and features that Google later adopted in its accused TPUs. For example:

Singular Presentations Made to Google / Jeff Dean (2010-2014)	**Google Documents**
 APPLICATIONS / MARKETS June 2011 Joe Bates Singular Computing LLC MIT Media Lab, Visiting Scientist Carnegie Mellon CS Department, Adjunct Prof joe@singularcomputing.com SINGULAR COMPUTING Joe Bates Singular Computing LLC Cambridge, MA (past) MIT Media Lab Carnegie Mellon Computer Science Dept joe@singularcomputing.com MULTI-MILLION-CORE PROCESSORS AND THEIR APPLICATIONS Joe Bates Singular Computing LLC Cambridge, MA (past) MIT Media Lab Carnegie Mellon Computer Science Dept joe@singularcomputing.com September 2013 MANY-MILLION CORE PROCESSORS AND THEIR APPLICATIONS Joe Bates Singular Computing LLC Cambridge, MA (past) MIT Media Lab Carnegie Mellon Computer Science Dept joe@singularcomputing.com January 2014	Google Publication of TPUv2 (2017) and TPUv3 (2018)  Machine Learning for Systems and Systems for Machine Learning Jeff Dean Google Brain team g.co/brain Presenting the work of **many** people at Google
 (SINGULAR'S) APPROXIMATE COMPUTING A traditional massively parallel machine, with floating point arithmetic that is "99% correct" (e.g. 1.0 + 1.0 = 1.98 ... 2.02)  • Surprise #1: Arithmetic circuit can be <u>unexpectedly</u> small Standard FPU ~500,000 transistors standard deterministic digital cmos 	 "We started to look at what we could do for these kinds of <u>deep learning models that could be more computationally efficient</u> and there are two really nice properties that deep learning models have. First, they are very tolerant of reduced precision... you don't need 6 or 7 digits of precision like you would in floating point computations or even more in double computations… you can build hardware that is only designed to accelerate low precision linear algebra, you're golden, and that enables you then really tailor the hardware to do only that,"

Singular Presentations Made to Google / Jeff Dean (2010-2014)	**Google Documents**
 OTHER PROMISING DOMAINS (IN PROGRESS - INITIAL EVIDENCE) • Vision: segmenting smooth objects (weak features, Hartmut/Joe intuition) • Molecular dynamics, Protein folding (all-atom energy) • Genomics (eg Smith-Waterman dynamic programming) • Machine learning (neural nets, genetic algorithms with local crossover, local graphical models, simulated annealing) • Speech recognition (HMMs, many concurrent voice streams, Dragon CTO) • Neocortex sim (>human, faster than realtime, supercomputer $) Confidential Property of Singular Computing June 2011 19	 Machine Learning for Systems and Systems for Machine Learning Jeff Dean Google Brain team g.co/brain Presenting the work of many people at Google "Around the time of maybe 2011, 2012, when the Google Brain project that I co-founded was just getting started, we started to collaborate with . . . the speech recognition team [at Google] . . . and so we could tell that as speech recognition gets better people are going to use it more and more . . . and at the time, we had [sic] lots and lots of CPUs in our data center and if you look at how much computation that would be required if a hundred million of our users started to do that, that was actually kind of daunting and scary, we would have essentially double the computing footprint of Google just to support like a slightly better speech recognition model.
 Combine FPU with 200 words memory. Build 2D grid, l Host CPU Control Unit Secondary Storage (DRAM) PE mem+fpu PE PE PE In vo 4K 1M (Confidential Property of Singular Computing June 2011	  HBM 8 GB core scalar/vector units MXU 128x128 core scalar/vector units MXU 128x128 HBM 8 GB

89. Due to its monitoring of Singular's patents and applications, Google knew of the application for the '616 patent prior to the issuance of the patent on August 25, 2020. For example, Google's attorneys prepared and filed six petitions for *Inter Partes* Review ("IPR") of patents related to the '616 patent. In each of those petitions, Google identified numerous patents and applications related to the '616 patent, as well as application serial number 16/882,686 for the '616 patent.

90. Before identifying the application, Google reviewed application serial number 16/882,686.

91. Google has known and/or should have known of the ’616 patent since its issuance or, at the latest, on or before Google’s receipt of service of this Complaint. Nonetheless, Google failed to cease its infringing activities or to seek a license to practice the inventions claimed in the patent. Alternatively, Google was willfully blind to application serial number 16/882,686 and to the issuance and its infringement of the ’616 patent.

92. As set forth above, in an attempt to prevent Google from stealing Dr. Bates’s inventions, Singular entered into an NDA with Google prior to Dr. Bates disclosing his inventions to representatives of Google. Notwithstanding the existence of the NDA, Google copied Dr. Bates’s inventions. Since the issuance of the ’616 patent, Google has done nothing to avoid infringing the ’616 patent.

93. Further evidence of Google’s nefarious conduct pertinent to this matter is Google’s attempt to induce Dr. Bates into executing the MNDA in 2017. The MNDA included a provision, drafted by counsel for Google, whereby Dr. Bates would, *inter alia*, waive claims for willful infringement. At the time that Google attempted to induce Dr. Bates into executing the MNDA, Google knew or should have known that it had incorporated Dr. Bates’s inventions previously disclosed by Dr. Bates to Google’s representatives into the design of the TPU v2 and/or TPU v3 and that Singular had patents covering such inventions.

94. When Google learned, or should have learned, of the issuance of the ’616 patent, Google should have ceased all manufacture, use, offering for sale and selling of the TPU v2 and TPU v3 devices that Google knew or should have known infringe one or more claims of the ’616 patent. Google knew or should have known that there was and is a high probability that the TPU

v2 and TPU v3 devices infringe the ’616 patent. Alternatively, Google was willfully blind to such facts.

95. At the time that Google began using the TPU v4 device in the United States, Google knew or should have known that the accused TPUs incorporated one or more of Dr. Bates’s patented inventions claimed in the ’616 patent. Google knew or should have known that there was and is a high probability that the TPU v4 devices infringe the ’616 patent. Alternatively, Google was willfully blind to such facts.

96. By the date of issuance of the ’616 patent, due to its knowledge of Singular’s Infringement Contentions served in case No. 1:19-cv-12551 (D. Mass.) involving patents related to the ’616 patent, Google knew or should have known that the accused TPUs infringe one or more claims of the ’616 patent and/or that there was a high probability that the accused TPUs infringe the ’616 patent. Alternatively, Google was willfully blind to such facts.

97. In view of, *inter alia*, its: (1) involvement in the ongoing patent litigation with Singular in this Court, including its knowledge of Singular’s Infringement Contentions therein; (2) IPR petitions regarding patent applications and patents owned by Singular; and (3) close monitoring of Singular’s patent portfolio, Google knew or should have known since at the latest on or around August 25, 2020 that there was a high risk that the accused TPUs infringe one or more claims of the ’616 patent. Alternatively, Google was willfully blind to the fact that such devices directly and/or indirectly infringe one or more claims of the ’616 patent.

98. Google’s actions described herein regarding the accused TPUs constitute conduct that was and continues to be willful, wanton, malicious, in bad-faith, deliberate, consciously wrong, flagrant and/or characteristic of a pirate. Google’s egregious conduct has continued unabated since the issuance of the ’616 patent.

99. Google's infringement of the '616 patent and willful conduct described above will continue unless and until Google is enjoined.

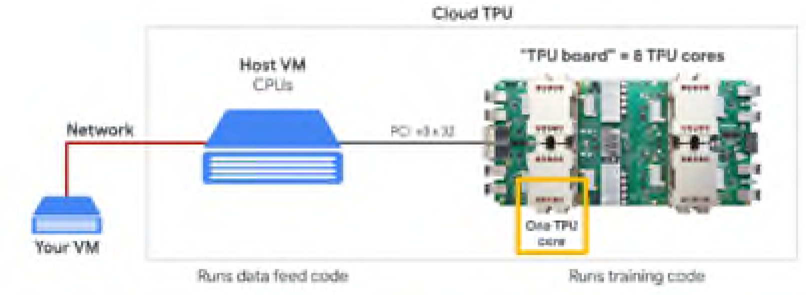
100. As a result of Google's infringement of the '616 patent, including its egregious and willful conduct described above, Singular has been irreparably harmed and suffered damages in an amount to be determined at trial.

COUNT II
INFRINGEMENT OF THE '775 PATENT

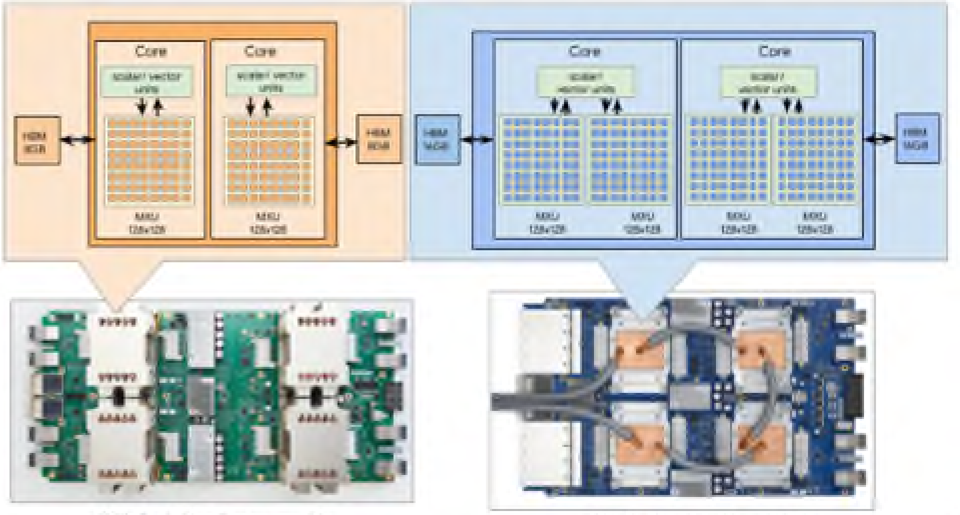
101. Paragraphs 1-100 above are incorporated herein by reference.

102. Google has directly infringed, and continues to directly infringe, literally and/or by the doctrine of equivalents, at least claims 1-2 and 20 of the '775 patent by making, testing, using, offering for sale, selling and/or importing into the United States the accused TPUs.

103. The accused TPUs comprise circuit boards that are connected to a host input-output CPU:



104. The accused TPUs comprise a plurality of MXU processing element arrays:



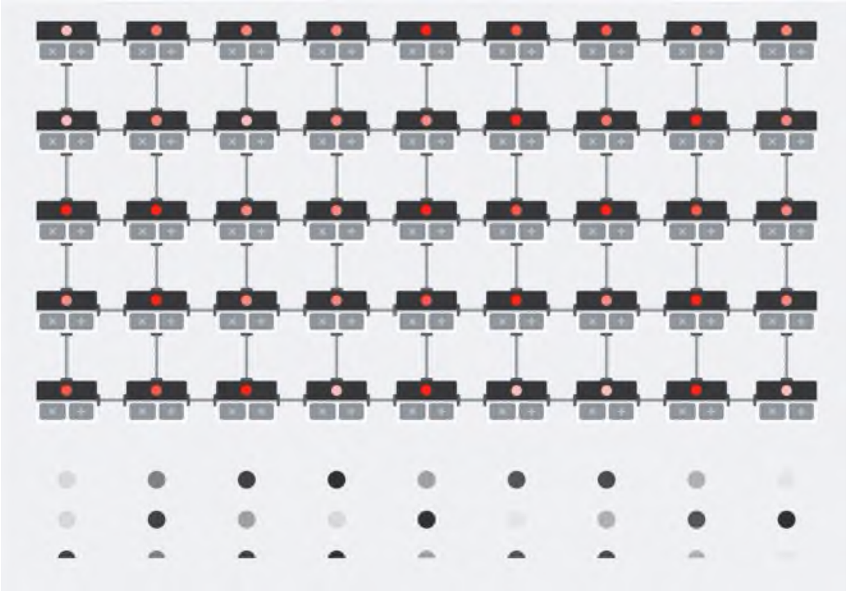
TPU v2 - 4 chips, 2 cores per chip

TPU v3 - 4 chips, 2 cores per chip

105. As shown above, the MXU processing elements are arranged in a 128x128 grid wherein at least one plurality of elements is positioned at a left edge of the array and a second plurality of elements is positioned in the interior of the array.

106. Each of the 128x128 MXUs in the accused TPUs comprises more than 5,000 processing elements. *See* https://cloud.google.com/tpu/docs/beginners-guide.

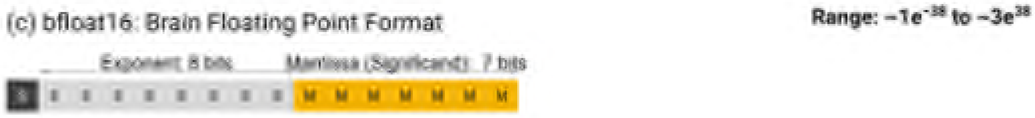
107. According to Google, the MXU processing elements are each connected to at least one other MXU processing element:



108. As shown above, each processing element comprises an arithmetic unit that performs multiplication.

109. Each of the MXU processing elements has an associated memory. This memory is used, for example, to store "weights" or "parameters" as part of algorithms that relate to neural networks. *See*, *e.g.*, https://cloud.google.com/tpu/docs/beginners-guide ("the TPU loads the parameters from memory into the matrix of multipliers and adders").

110. When performing multiplication using bfloat16, floating point values having a 7-bit mantissa (plus a sign bit) and an 8-bit binary exponent are input into the arithmetic units:


(c) bfloat16: Brain Floating Point Format
Range: ~1e-38 to ~3e38
Exponent: 8 bits
Mantissa (Significand): 7 bits
S
E E E E E E E E
M M M M M M M

111. Each Scalar Vector Unit ("SVU") (a/k/a Vector Processing Unit) in the accused TPUs comprises multipliers that receive 32-bit wide inputs. *See*, *e.g.*, https://codelabs.developers.google.com/codelabs/keras-flowers-data/#2 ("The VPU handles float32 and int32 computations").

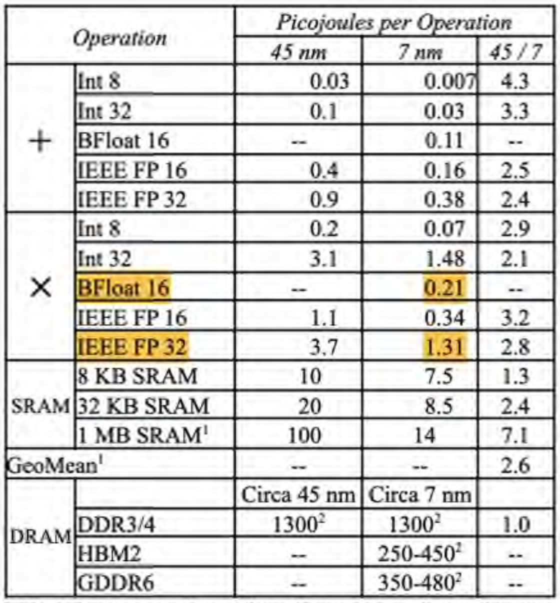
112. Each MXU has 128x128 = 16,384 processing elements that operate on bfloat16 numbers. The accused TPUs have at least 2 MXUs each, for a total of at least 32,768 processing elements. *See*, *e.g.*, https://cloud.google.com/tpu/docs/beginners-guide. This is at least 20 more than three times the number of 32-bit multipliers in the accused TPUs.

113. Each bfloat16 multiplier within the MXU comprises fewer transistors than each 32-bit multiplier within the SVU. Google engineer Jeffrey Dean, the head of Google Brain, expressly admitted this:

> Furthermore, one major area & power cost of multiplier circuits for a floating point format with M mantissa bits is the (*M*+1) ✕ (M+1) array of full adders (that are needed for multiplying together the mantissa portions

> of the two input numbers. The IEEE fp32, IEEE fp16 and bfloat16 formats need ***576 full adders***, 121 full adders, and ***64 full adders***, respectively. ***Because multipliers for the bfloat16 format require so much less circuitry***, it is possible to put more multipliers in the same chip area and power budget, thereby meaning that ML accelerators employing this format can have higher flops/sec and flops/Watt, all other things being equal.

Dean, Jeffrey. (2020). 1.1 *The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design*. 8-14. 10.1109/ISSCC19947.2020.9063049 (emphasis added). This fact was further confirmed in a paper published by the team of Google engineers responsible for designing and building the accused TPUs (including, *inter alia*, Norman Jouppi and David Patterson):



Operation		*Picojoules per Operation*		
		45 nm	*7 nm*	*45 / 7*
+	Int 8	0.03	0.007	4.3
	Int 32	0.1	0.03	3.3
	BFloat 16	--	0.11	--
	IEEE FP 16	0.4	0.16	2.5
	IEEE FP 32	0.9	0.38	2.4
×	Int 8	0.2	0.07	2.9
	Int 32	3.1	1.48	2.1
	BFloat 16	--	0.21	--
	IEEE FP 16	1.1	0.34	3.2
	IEEE FP 32	3.7	1.31	2.8
SRAM	8 KB SRAM	10	7.5	1.3
	32 KB SRAM	20	8.5	2.4
	1 MB SRAM[1]	100	14	7.1
GeoMean[1]		--	--	2.6
DRAM		Circa 45 nm	Circa 7 nm	
	DDR3/4	1300[2]	1300[2]	1.0
	HBM2	--	250-450[2]	--
	GDDR6	--	350-480[2]	--

Table 2. Energy per Operation: 45 nm [16] vs 7 nm. Memory is pJ per 64-bit access.

Jouppi, Norman, *et al.*. "Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, Valencia, Spain, 2021 pp. 1-14 at 3 (emphasis added). According to the above table, a bfloat16 multiplier requires less than 20% as much energy per operation as an IEEE FP32 multiplier (both made using the same 7 nm semiconductor fabrication process). The lower power requirements of bfloat16 multipliers is a result of the fact that they include fewer transistors than full-precision IEEE FP32 multipliers.

114. As a result of Google's IPR petitions and activities described above, including its monitoring of Singular's patent applications and patents, Google knew of application serial number 16,882,694 which led to the '775 patent since, at the latest, October 30, 2020 when Google identified the application in, *inter alia*, its Petition for *Inter Partes* Review in IPR2021-00154. Before making such identification, counsel for Google reviewed application serial number 16/882,694.

115. Google's infringement of the '775 patent will continue unless and until Google is enjoined.

116. As a result of Google's infringement of the '775 patent, Singular has been irreparably harmed and suffered damages in an amount to be determined at trial.

PRAYER FOR RELIEF

WHEREFORE, Singular requests that the Court:

A. enter judgment in favor of Singular on both counts of the complaint;

B. award Singular damages resulting from Google's past and ongoing infringing conduct, together with interest thereon and costs pursuant to 35 U.S.C. § 284;

C. award Singular enhanced damages pursuant to 35 U.S.C. § 284;

D. award Singular its attorney's fees incurred herein pursuant to 35 U.S.C. § 285;

E. enjoin Google's infringement of the '616 patent and the '775 patent pursuant to 35 U.S.C. § 283, and

F. award Singular such other and further legal and equitable relief as the Court may deem just and proper.

DEMAND FOR JURY TRIAL

Singular demands a trial by jury on all issues so triable.

Dated: December 22, 2021

Respectfully submitted,

/s/ Paul J. Hayes
Paul J. Hayes (BBO #227000)
Matthew D. Vella (BBO #660171)
Kevin Gannon (BBO #640931)
Brian M. Seeve (BBO #670455)
Daniel McGonagle (BBO #690084)
Michael J. Ercolini (*pro hac vice*)
Thomas R. Fulford (BBO #686160)
PRINCE LOBEL TYE LLP
One International Place, Suite 3700
Boston, MA 02110
Tel: (617) 456-8000
Email: phayes@princelobel.com
Email: mvella@princelobel.com
Email: kgannon@princelobel.com
Email: bseeve@pricelobel.com
Email: dmcgonagle@princelobel.com
Email: mercolini@princelobel.com
Email: tfulford@princelobel.com

ATTORNEYS FOR THE PLAINTIFF